数学基础班第 2 课课件：微积分选讲

管枫

七月在线

July, 2016

主要内容

- 极限
  - 复习极限记号，无穷大无穷小阶数
- 微分学
  - 复习函数求导，泰勒级数逼近
  - 牛顿法与梯度下降法
- Jensen 不等式
  - 复习凸函数，Jensen 不等式的证明

记号

- 本节课常用数学记号

$\mathbb{R}^n$	实坐标空间
$f(x)$	函数
$\lim\limits_{x \to x_0} f(x)$	函数的极限
$O(x^n)$	当 x 趋于 0 时的 n 阶无穷小
$o(x^n)$	当 x 趋于 0 时的 n 阶以上无穷小
f'	函数导数
$\frac{df}{dx}$	函数导数
$\frac{\partial f}{\partial x}$	函数在 x 坐标方向上的偏导数
$\nabla_v f$	函数在 v 方向上的方向导数
$\int_a^b f(x)dx$	函数积分

极限

通俗语言

函数 f 在 x_0 处的极限为 L

数学记号

$$\lim_{x \to x_0} f(x) = L$$

精确描述:$\epsilon - \delta$ 语言

对于任意的正数 $\epsilon > 0$, 存在正数 δ, 使得任何满足 $|x - x_0| < \delta$ 的 x, 都有

$$|f(x) - L| < \epsilon$$

通俗语言适合于说给对方听，数学记号适合于写给对方看，精确描述比较啰嗦但是非常精确不会造成误解，主要用于证明.

极限: 如何比较无穷小

无穷也分大小，如何描述与比较无穷大和无穷小

Example

当 x 趋于 0 的时候，$\sin(x)$ 与 $\tan(x)$ 都趋于 0. 但是哪一个趋于 0 的速度更快一些呢?

我们考察这两个函数的商的极限

$$\lim_{x\to 0}\frac{\sin(x)}{\tan(x)} = \lim_{x\to 0}\cos(x) = \cos(0) = 1$$

所以当 $x \to 0$ 的时候，$\sin(x)$ 与 $\tan(x)$ 是同样级别的无穷小.

极限: 无穷小阶数

Definition (无穷小阶数)

当 $x \to 0$ 时,

- 如果 $\lim\limits_{x\to 0} f(x) = 0$ 而且 $\lim\limits_{x\to 0} f(x)/x^n = 0$ 那么此时 $f(x)$ 为 n 阶以上无穷小，记为

$$f(x) = o(x^n), x \to 0$$

- 如果 $\lim\limits_{x\to 0} f(x) = 0$ 而且 $\lim\limits_{x\to 0} f(x)/x^n$ 存在且不等于零，那么此时 $f(x)$ 为 n 阶无穷小，记为

$$f(x) = O(x^n), x \to 0$$

为了方便，在不至于引起误解的时候我们回省略掉 $x \to 0$.

所谓无穷小的阶数，就是用我们比较熟悉的多项式类型的无穷小量来衡量其他的无穷小量. 把未解决的问题转化为已经解决的问题，是数学家的惯用伎俩.

极限: 如何比较无穷小

Proposition (三明治/两边夹/夹逼原理)

如果三个函数满足 $f(x) \leq g(x) \leq h(x)$, 而且他们都在 x_0 处有极限，那么

$$\lim_{x \to x_0} f(x) \leq \lim_{x \to x_0} g(x) \leq \lim_{x \to x_0} h(x)$$

极限: 如何比较无穷小

Example (重要极限)

- $\lim_{x\to 0} \sin(x)/x = 1$
- $\lim_{x\to \infty} x^{\alpha}/e^{x} = 0$, 对于任意正数 α
- $\lim_{x\to \infty} \ln(x)/x^{\alpha} = 0$, 对于任意正数 α
- $\lim_{x\to \infty} (1+1/x)^{x} = e$

小结 (极限)

- 极限有不同的表述方式，建议练习使用 $\epsilon - \delta$ 语言
- 无穷小也可以互相比较
- 利用多项式类型无穷小，可以定义无穷小的阶数
- 夹逼原理

微分学的核心思想: 逼近.

Definition (函数的导数)

如果一个函数 $f(x)$ 在 x_0 附近有定义，而且存在极限

$$L = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}.$$

那么 $f(x)$ 在 x_0 处可导且导数 $f'(x_0) = L$.

等价定义

无穷小量表述: 线性逼近

如果存在一个实数 L 使得 $f(x)$ 满足,

$$f(x) = f(x_0) + L(x - x_0) + o(x - x_0), x \to x_0.$$

那么 $f(x)$ 在 x_0 处可导且导数 $f'(x_0) = L$.

Definition (函数的高阶导数)

如果函数的导数函数仍然可导，那么导数函数的导数是二阶导数，二阶导数函数的导数是三阶导数. 一般地记为

$$f^{(n)}(x) = \frac{d}{dx} f^{(n-1)}(x)$$

或者进一步

$$f^{(n)}(x) = \frac{d^n}{dx^n} f(x)$$

导数是对函数进行线性逼近，高阶导数是对导数函数的进一步逼近，因为没有更好的办法，所以数学家选择继续使用线性逼近.

Example (初等函数的导数)

$$\frac{d}{dx}\sin(x) = \cos(x) \qquad \frac{d}{dx}\cos(x) = -\sin(x)$$

$$\frac{d}{dx}\sinh(x) = \cosh(x) \qquad \frac{d}{dx}\cosh(x) = \sinh(x)$$

$$\frac{d}{dx}x^n = nx^{n-1} \qquad \frac{d^n}{dx^n}x^n = n!$$

$$\frac{d}{dx}e^x = e^x \qquad \frac{d}{dx}\ln(x) = 1/x$$

我们在此只考虑无穷次可微的多元函数，对于这种函数我们可以使用全微分来定义偏导数，实际上现实遇到的函数大部分都是这种函数.

Definition (全微分与偏导数)

以二元函数为例, 如果 $f(x,y)$ 是一个二元函数，而且存在 L_x 和 L_y 使得:

$$f(x_0+\Delta_x, y_0+\Delta_y) = f(x_0,y_0) + L_x\Delta_x + L_y\Delta_y + o(|\Delta_x|+|\Delta_y|)$$

那么 $f(x,y)$ 在 (x_0,y_0) 点处可微

且 L_x, L_y 分别是 f 在 x,y 方向上的偏导数. 一般记为

$$\frac{\partial}{\partial x}f(x_0,y_0) = L_x \quad \frac{\partial}{\partial y}f(x_0,y_0) = L_y$$

Definition (高阶偏导数)

以二元函数为例, 如果 $f(x,y)$ 是一个二元函数，而且存在 $L_x, L_y, L_{xy}, L_{x^2}, L_{y^2}$ 使得:

$$\begin{aligned} f(x_0+\Delta_x, y_0+\Delta_y) =& f(x_0,y_0) + L_x\Delta_x + L_y\Delta_y \\ &+ L_{xy}\Delta_x\Delta_y + \frac{L_{x^2}}{2}\Delta_x^2 + \frac{L_{y^2}}{2}\Delta_y^2 \\ &+ o(|\Delta_x|^2 + |\Delta_y|^2) \end{aligned}$$

那么 $f(x,y)$ 在 (x_0,y_0) 点处二阶可微

并且二阶偏导数为

$$\frac{\partial^2}{\partial x^2} f(x_0,y_0) = L_{x^2}, \quad \frac{\partial^2}{\partial y^2} f(x_0,y_0) = L_{y^2}, \quad \frac{\partial^2}{\partial y \partial x} f(x_0,y_0) = L_{xy}$$

Example (偏导数的例子)

$f(x,y) = x^2 + 2xy + 3y^2$, 则

$$\frac{\partial}{\partial x} f(x,y) = 2x + 2y, \quad \frac{\partial}{\partial y} f(x,y) = 2x + 6y$$

$$\frac{\partial^2}{\partial x^2} f(x,y) = 2, \quad \frac{\partial^2}{\partial y^2} f(x,y) = 6, \quad \frac{\partial^2}{\partial x \partial y} f(x,y) = 2$$

Example (偏导数的例子)

$f(x,y)=\ln(x+y^2)$, 则

$$\frac{\partial}{\partial x}f(x,y)=\frac{1}{x+y^2},\quad \frac{\partial}{\partial y}f(x,y)=\frac{2y}{x+y^2}$$

求导法则

- 链式法则: $\frac{d}{dx}(g \circ f) = \frac{dg}{dx}(f) \cdot \frac{df}{dx}$
- 加法法则: $\frac{d}{dx}(g + f) = \frac{dg}{dx} + \frac{df}{dx}$
- 乘法法则: $\frac{d}{dx}(g \cdot f) = \frac{dg}{dx} \cdot f + g \cdot \frac{df}{dx}$
- 除法法则: $\frac{d}{dx}(\frac{g}{f}) = \frac{\frac{dg}{dx} \cdot f - \frac{df}{dx} \cdot g}{f^2}$
- 反函数求导: $\frac{d}{dx}(f^{-1}) = \frac{1}{\frac{df}{dx}(f^{-1})}$

所有求导法则原则上都可以由链式法则结合二元函数的偏导数来推出来，有兴趣的同学可以思考一下这是为什么

Example (例子)

$f(x) = x^x$, 求导数函数

首先简化一下 $f(x) = x^x = \exp(\ln(x))^x = \exp(\ln(x) \cdot x)$. 于是问题转化为了一个复合函数的样子, 具体说来如果
$g(x) = \exp(x)$,$h(x) = x\ln(x)$, 则 $f(x) = (g \circ h)(x)$

$$\begin{aligned} f'(x) &= g'(h(x)) \cdot h'(x) = g(h(x)) \cdot h'(x) \\ &= f(x) \cdot (\ln(x) + x \cdot \frac{1}{x}) \\ &= f(x) \cdot (\ln(x) + 1) \\ &= x^x(\ln(x) + 1) \end{aligned}$$

小结 (求导)

- 微分学的核心思想是逼近.
- 一阶导数：线性逼近
- 二阶导数：二次逼近
- 导数计算：求导法则

一元微分学的顶峰：泰勒级数

用多项式逼近的方式描述高阶导数，我们就得到了泰勒级数.

泰勒/迈克劳林级数: 多项式逼近

如果 $f(x)$ 是一个无限次可导的函数，那么在任何一点 x_0 附近我们可以对 $f(x)$ 做多项式逼近:

$$\begin{aligned} f(x_0+\Delta_x) =& f(x_0)+f'(x_0)\Delta_x+\frac{f''(x_0)}{2}\Delta_x^2+\cdots \\ &+\frac{f^{(n)}(x_0)}{n!}\Delta_x^n+o(\Delta_x^n) \end{aligned}$$

在本课中我们不关注对于尾巴上的余项 $o(\Delta_x^n)$ 的大小估计 常庚哲和史济怀老师把泰勒级数称为是一元微分学的顶峰，我也同意这个观点

一元微分学的顶峰：泰勒级数

Example

泰勒级数: 例子

- $e^x = 1 + x + x^2/2 + \cdots + x^n/n! + o(x^n)$
- $\ln(1+x) = x - x^2/2 + x^3/3 + \cdot + (-1)^{n-1}x^n/n + o(x^n)$
- $\sin(x) = x - x^3/6 + \cdots + (-1)^n x^{2n+1}/(2n+1)! + o(x^{2n+1})$
- $\cos(x) = 1 - x^2/2 + x^4/24 + \cdots + (-1)^n x^{2n}/(2n)! + o(x^{2n})$

有兴趣的同学可以思考一下为什么欧拉能写下那个号称史上最美的数学公式 $e^{i\theta} = \cos(\theta) + i\sin(\theta)$.

用泰勒级数来理解问题

泰勒级数是一元微分逼近的顶峰，所以有关于一元微分逼近的问题请尽情使用.

罗比塔法则

如果 f, g 是两个无穷阶可导的函数，而且 $f(x_0) = g(x_0) = 0$, $g'(x_0) \neq 0$, 则 $\lim\limits_{x \to x_0} f(x)/g(x) = \lim\limits_{x \to x_0} f'(x)/g'(x)$.

因为是在 x_0 附近的极限问题，我们使用泰勒级数来思考这个问题

$$f(x_0 + \Delta_x) = f(x_0) + f'(x_0)\Delta_x + o(\Delta_x)$$
$$g(x_0 + \Delta_x) = g(x_0) + g'(x_0)\Delta_x + o(\Delta_x)$$

用泰勒级数来理解问题

$$\begin{aligned}
\lim_{x\to x_0} f(x)/g(x) &= \lim_{x\to x_0} \frac{f(x_0)+f'(x_0)\Delta_x+o(\Delta_x)}{g(x_0)+g'(x_0)\Delta_x+o(\Delta_x)} \\
&= \lim_{\Delta_x\to 0} \frac{f'(x_0)\Delta_x+o(\Delta_x)}{g'(x_0)\Delta_x+o(\Delta_x)} \\
&= \lim_{\Delta_x\to 0} \frac{f'(x_0)+o(\Delta_x)/\Delta_x}{g'(x_0)+o(\Delta_x)/\Delta_x} \\
&= \frac{f'(x_0)}{g'(x_0)}
\end{aligned}$$

用泰勒级数来理解问题

求解简单的微分方程

求解满足如下方程组的解析函数 $f(x)$

$$f''(x) = -f(x), \quad f(0) = 0, \quad f'(0) = 1$$

我们没有讲过微分方程的概念，但是这个问题已经可以解决了. 因为 f 是一个光滑函数那么它在 0 点附近就有泰勒级数

$$f(x) = f(0) + f'(0)x + f''(0)x^2/2 + f^{(3)}(0)x^3/6 + \cdots$$

或者写成

$$f(x) = \sum_{n=0}^{\infty} f^{(n)}(0)x^n/n!$$

，于是

$$f''(x) = \sum_{n=0}^{\infty} f^{(n+2)}(0)x^n/n!$$

用泰勒级数来理解问题

于是第一个方程给出了递推公式

$$f^{(n+2)}(0) = -f^{(n)}(0)$$

第二个方程和第三个方程给出了初始值

$$f(0) = 0, \quad f'(0) = 1$$

, 于是

$$f^{(2n)} = 0, \quad f^{(2n+1)} = (-1)^n$$

所以 $f(x)$ 的泰勒级数只能是

$$f(x) = x - x^3/6 + \cdots + (-1)^n x^{2n+1}/(2n+1)! + \cdots$$

对照之前的特殊函数泰勒级数我们知道: $f(x) = \sin(x)$.

小结 (泰勒级数)

- 泰勒级数本质是多项式逼近
- 特殊函数的泰勒级数可以适当记一下
- 泰勒级数可以应用于很多与逼近相关的问题

牛顿法与梯度下降法

很多机器学习或者统计的算法最后都转化成一个优化的问题. 也就是求某一个损失函数的极小值的问题, 在本课范围内我们考虑可微分的函数极小值问题.

优化问题

对于一个无穷可微的函数 $f(x)$，如何寻找他的极小值点.

极值点条件

- 全局极小值: 如果对于任何 $\tilde{x}$, 都有 $f(x_*) \leq f(\tilde{x})$，那么 x_* 就是全局极小值点.
- 局部极小值: 如果存在一个正数 δ 使得，对于任何满足 $|\tilde{x} - x_*| < \delta$ 的 $\tilde{x}$, 都有 $f(x_*) \leq f(\tilde{x})$，那么 x_* 就是局部极小值点. (方圆 δ 内的极小值点)
- 不论是全局极小值还是局部极小值一定满足一阶导数/梯度为零，$f' = 0$ 或者 $\nabla f = 0$.

牛顿法与梯度下降法

局部极值算法

我们本节课利用极值点条件，来介绍牛顿法和梯度下降法.

- 这两种方法都只能寻找局部极值
- 这两种方法都要求必须给出一个初始点 x_0
- 数学原理：牛顿法使用二阶逼近，梯度下降法使用一阶逼近
- 牛顿法对局部凸的函数找到极小值，对局部凹的函数找到极大值，对局部不凸不凹的可能会找到鞍点.
- 梯度下降法一般不会找到最大值，但是同样可能会找到鞍点.
- 当初始值选取合理的情况下，牛顿法比梯度下降法收敛速度块.
- 牛顿法要求估计二阶导数，计算难度更大.

牛顿法与梯度下降法

牛顿法：二次逼近

首先在初始点 x_0 处，写出二阶泰勒级数

$$f(x_0+\Delta_x) = f(x_0) + f'(x_0)\Delta_x + \frac{f''(x_0)}{2}\Delta_x^2 + o(\Delta_x^2) \tag{1}$$

$$= g(\Delta_x) + o(\Delta_x^2) \tag{2}$$

我们知道关于 Δ_x 的二次函数 $g(\Delta_x)$ 的极值点为 $-\frac{f'(x_0)}{f''(x_0)}$. 那么本着逼近的精神 $f(x)$ 的极值点估计在 $x_0 - \frac{f'(x_0)}{f''(x_0)}$ 附近，于是定义 $x_1 = x_0 - \frac{f'(x_0)}{f''(x_0)}$，并重复此步骤得到序列

$$x_n = x_{n-1} - \frac{f'(x_{n-1})}{f''(x_{n-1})}$$

当初始点选的比较好的时候 $\lim\limits_{n\to\infty} x_n$ 收敛于一个局部极值点.

牛顿法与梯度下降法

牛顿法：多变量函数二阶逼近

如果函数 $f(x)$ 是个多元函数,x 是一个向量，那么牛顿法序列变为:

$$x_n = x_{n-1} - (\mathbb{H}f(x_{n-1}))^{-1} \cdot \nabla f(x_{n-1})$$

思路与技巧完全相同，只是使用梯度 ∇f 取代一阶导数 f', 使用 Hessian 矩阵 $\mathbb{H}f$ 代替二阶导数 f''.

牛顿法与梯度下降法

梯度下降法：多变量函数一阶逼近

如果函数 $f(x)$ 是个多元函数,x 是一个向量. 在 x_0 处对 f 做线性逼近

$$\begin{aligned} f(x_0+\Delta_x) &= f(x_0)+\Delta_x^T\cdot\nabla f(x_0)+o(|\Delta_x|) \\ &= g(\Delta_x)+o(|\Delta_x|) \end{aligned}$$

但是线性函数 $g(\Delta_x)$ 是没有极值点的！所以这个线性逼近不能告诉我们极值点在什么地方，他只能告诉我们极值点在什么方向. 所以我们只能选取一个比较“小”的 γ 沿着这个方向走下去,并得到梯度下降法的序列:

$$x_n = x_{n-1}-\gamma_{n-1}\nabla f(x_{n-1})$$

牛顿法与梯度下降法

小结 (牛顿法与梯度下降法)

- 牛顿法与梯度下降法本质上都是对目标函数进行局部逼近.
- 因为是局部逼近所以也只能寻找局部极值
- 牛顿法收敛步骤比较少，但是梯度下降法每一步计算更加简单
- 不同的算法之间很难说哪一个更好，选择算法还要具体问题具体分析（这也是数据科学家存在的意义之一）

凸函数与琴生不等式

为什么研究凸函数，凸优化?

对于凸优化来说，局部极值与整体极值没有区别. 这个比较简单，所以从这个开始研究.

当不知道该做什么的时候，选择最简单的问题开始做起，这也是数学家的惯用伎俩.

凸函数与琴生不等式

Definition (凸函数)

一个函数 f 如果满足

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2), \forall \lambda \in (0,1)$$

那么这个函数就是凸函数.

把如上定义中的 $\leq$ 换成 $<$, 那么这个函数就叫做严格凸函数.

凸函数与琴生不等式

Proposition (凸函数判断准则)

一个函数二阶可微的函数 f 是凸函数，当且仅当 $f''(x) \geq 0, \forall x.$

如果 f 是多元函数,x 是个向量, 那么 f 是凸函数的条件变为 $\mathbb{H}f$ 是一个半正定矩阵.

凸函数与琴生不等式

Proposition (凸函数重要性质: 琴生不等式)

如果 f 是凸函数，那么对于任意的 $\{x_1, x_2, \cdots, x_n\}$, 以及正的权重系数 $\{w_1, w_2, \cdots, w_n\}$, 且 $w_1 + w_2 + \cdots + w_n = 1$, 则如下不等式成立

$$f(\sum_{k=1}^{n} w_k \cdot x_k) \leq \sum_{k=1}^{n} w_k \cdot f(x_k)$$

根据定义我们已经知道当 $n = 2$ 的时候此结论成立，那么对于一般的 n 我们采取数学归纳法来进行证明.

Proof

假设此结论对于 $n = N$ 成立，我们考虑当 $n = N + 1$ 的情形. 此时我们有 $N + 1$ 个点 $x_1, ..., x_{N+1}$，以及权重系数 $w_1, ..., w_{N+1}$. 令 $S_N = \sum_{k=1}^{N} w_k$，$A_N = \sum_{k=1}^{N} w_k x_k$，并定义 $u_k = w_k / S_N, \forall k \in \{1, .., N\}$. 那么根据归纳假设我们有

$$f(A_N/S_N) = f(\sum_{k=1}^{N} u_k \cdot x_k) \leq \sum_{k=1}^{N} u_k \cdot f(x_k) = \frac{\sum_{k=1}^{N} w_k \cdot f(x_k)}{S_N}$$

Continue Proof

于是得到

$$S_N f(A_N/S_N) \leq \sum_{k=1}^{N} w_k \cdot f(x_k)$$

两边同时加上 $w_{N+1}f(x_{N+1})$, 得到

$$S_N f(A_N/S_N) + w_{N+1}f(x_{N+1}) \leq \sum_{k=1}^{N+1} w_k \cdot f(x_k) \tag{3}$$

Continue Proof

而另一方面根据凸函数的定义，由于 $S_N + w_{N+1} = 1$, 我们有

$$f(\sum_{k=1}^{N+1} w_k \cdot x_k) = f(S_N \cdot A_N/S_N + w_{N+1}x_{N+1}) \tag{4}$$

$$\leq S_N f(A_N/S_N) + w_{N+1} f(x_{N+1}) \tag{5}$$

不等式(3)与不等式(5)合起来即得所证.

从 $n = N$ 的情形出发，证明了 $n = N + 1$ 的情形。而且 $n = 2$ 的情形已知成立，由数学归纳法，原定理证明完毕.
(采用数学归纳法，从简单到复杂的研究问题，也是数学家的惯用伎俩.)

谢谢大家!